IDF Revisited: A Simple New Derivation within the Robertson-Spärck Jones Probabilistic Model

Lillian Lee
Dept. of Computer Science, Cornell University
Ithaca, NY 14853-7501 USA
http://www.cs.cornell.edu/home/llee
llee@cs.cornell.edu

ABSTRACT

There have been a number of prior attempts to theoretically justify the effectiveness of the inverse document frequency (IDF). Those that take as their starting point Robertson and Spärck Jones's probabilistic model are based on strong or complex assumptions. We show that a more intuitively plausible assumption suffices. Moreover, the new assumption, while conceptually very simple, provides a solution to an estimation problem that had been deemed intractable by Robertson and Walker (1997).

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms: Theory, Algorithms

Keywords: inverse document frequency, IDF, probabilistic model, term weighting

1. INTRODUCTION

The inverse document frequency (IDF) [12] has been "incorporated in (probably) all information retrieval systems" ([6], pg. 77). Attempts to theoretically explain its empirical successes abound ([2, 14, 1, 11, 5, 8, 4, 3], *inter alia*). Our focus here is on explanations based on Robertson and Spärck Jones's *probabilistic-model* (RSJ-PM) paradigm of information retrieval [10], not because of any prejudice against other paradigms, but because a certain RSJ-PM-based justification of the IDF in the absence of relevance information has been promulgated by several influential authors [2, 9, 7].

RSJ-PM-based accounts use either an assumption due to Croft and Harper [2] that is mathematically convenient but not plausible in real settings, or a complex assumption due to Robertson and Walker [11]. We show that the IDF can be derived within the RSJ-PM framework via a new assumption that directly instantiates a highly intuitive notion, and that, while conceptually simple, solves an estimation problem deemed intractable by Robertson and Walker [11].

2. CROFT-HARPER DERIVATION

In the (binary-independence version of the) RSJ-PM, the i^{th} term is assigned weight

$$\log \frac{p_i(1-q_i)}{q_i(1-p_i)}\,. \tag{1}$$

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

where $p_i \stackrel{\text{def}}{=} P(X_i = 1 \,|\, R = \mathtt{y})$, $q_i \stackrel{\text{def}}{=} P(X_i = 1 \,|\, R = \mathtt{n})$, X_i is an indicator variable for the presence of the i^{th} term, and R is a relevance random variable. Croft and Harper [2] proposed the use of two assumptions to estimate p_i and q_i in the absence of relevance information. **CH-1**, which is unobjectionable, simply states that most of the documents in the corpus are not relevant to the query. This allows us to set $\widehat{q_i^{\text{CH}}} \stackrel{\text{def}}{=} \frac{n_i}{N}$, where n_i is the number of documents in the corpus that contain the i^{th} term, and N is the number of documents in the corpus. The second assumption, **CH-2**, is that all query terms share the same probability π of occurring in a relevant document[1]. Under CH-2, one sets $\widehat{p_i^{\text{CH}}} \stackrel{\text{def}}{=} \pi$, and thus (1) becomes

$$\pi' + \log \frac{N-n_i}{n_i}\,, \tag{2}$$

where $\pi' = \log(\pi/(1-\pi))$ is constant (and is 0 if $\pi = .5$). Quantity (2) is essentially the IDF.

CH-2 is an ingenious device for pushing the derivation above through. However, intuition suggests that the occurrence probability of query terms in relevant documents should be at least somewhat correlated with their occurrence probability in arbitrary documents within the corpus, and hence not constant. For example, a very frequent term can be expected to occur in a noticeably large fraction of any particular subset of the corpus, including the relevant documents. Contrariwise, a query term might be relatively infrequent overall due to having a more commonly used synonym; such a term would still occur relatively infrequently even within the set of (truly) relevant documents.[2]

3. ROBERTSON-WALKER DERIVATION

Robertson and Walker (RW) [11] also object to CH-2, on the grounds that for query terms with very large document frequencies, weight (2) can be negative. This anomaly, they show, arises precisely because $\widehat{p_i^{\text{CH}}}$ is constant. They then propose the following alternative:

$$\widehat{p_i^{\text{RW}}} \stackrel{\text{def}}{=} \frac{\pi}{\pi + (1-\pi)\frac{N-n_i}{N}}\,,$$

where π is the Croft-Harper constant, but reinterpreted as the estimate for p_i just when $n_i = 0$. One can check that $\widehat{p_i^{\text{RW}}} \in [\pi, 1]$ slopes up hyperbolically in n_i. Applying $\widehat{p_i^{\text{RW}}}$

[1]This can be relaxed to apply to just the query terms in the document in question.
[2]Indeed, one study [5] did find p_i increasing with n_i.

and $\widehat{q_i^{\text{CH}}}$ to the term-weight scheme (1) yields

$$\pi' + \log \frac{N}{n_i} \tag{3}$$

(which is positive as long as $\pi \geq .5$).

4. NEW ASSUMPTION

The estimate $\widehat{p_i^{\text{RW}}}$ increases monotonically in n_i, which is a desirable property, as we have argued above. However, its exact functional form does not seem particularly intuitive. RW motivate it simply as an approximation to a linear form; approximation is necessary, they claim, because

> the straight-line model [i.e., p_i linear in q_i and hence n_i by CH-1] is actually rather intractable, and does not lead to a simple weighting formula ([11], pg. 18).[3]

Despite this claim, we show here that there exists a highly intuitive linear estimate that leads to a term weight varying inversely with document frequency.

There are two main principles that motivate our new estimate. First, as already stated, any estimate of p_i should be positively correlated with n_i. The second and key insight is that *query terms should have a higher occurrence probability within relevant documents than within the document collection as a whole*. Thus, if the i^{th} term appears in the query, we should "lift" its estimated occurrence probability in relevant documents above n_i/N, which is its estimated occurrence probability in general documents. This leads us to the following intuitive estimate, which is reminiscent of "add-one smoothing" used in language modeling (more on this below):

$$\widehat{p_i} \stackrel{\text{def}}{=} \frac{n_i + L}{N + L}. \tag{4}$$

Here the $L > 0$ in the numerator[4] is a "lift" or "boost" constant.[5] Plugging $\widehat{p_i}$ and $\widehat{q_i^{\text{CH}}}$ into (1) yields the term weight

$$\log \left(\frac{\frac{n_i+L}{N+L}}{\frac{n_i}{N}} \frac{\frac{N-n_i}{N}}{\frac{N-n_i}{N+L}} \right) = \log \left(1 + \frac{L}{n_i} \right),$$

which varies inversely in n_i, as desired.

Furthermore, as hinted at above, selecting L's value is equivalent to selecting $\widehat{p_i}$'s value for query terms whose document frequency is 0. That is, $L/(N + L)$ is directly analogous to π in RW's derivation. Indeed, choosing $L = N$ is just like choosing $\pi = 0.5$, which is commonly done in presentations of the Croft-Harper derivation in order to eliminate the leading constant π' in (2); doing so in our case yields the following term weight, which is the "usual" form of the IDF ([13], pg. 184):

$$\log \left(1 + \frac{N}{n_i} \right).$$

[3]The fact that RW's Figure 2 depicts the linear scenario graphically appears to have led to some mistaken impressions (e.g., [5], pg. 18, coincidentally) that this is the mathematical model that RW actually employed.

[4]The L in the denominator ensures that $\widehat{p_i} \leq 1$.

[5]Since the RSJ-PM document-scoring function only accumulates weights for terms appearing in the query, it is fine to compute the $\widehat{p_i}$'s offline, that is, before the query is seen.

Finally, note that $\widehat{p_i}$ is linear in n_i; we have thus contradicted the assertion quoted above that developing a "straight-line" model is "intractable" [11].

5. ONWARD AND UPWARD

An interesting direction for future work is to consider lift *functions* $L(n_i)$ that depend on n_i. It can be shown that different choices of $L(n_i)$ allow one to model *non-linear* dependencies of p_i on n_i that occur in real data, such as the approximately logarithmic dependence observed in TREC corpora by Greiff [5]. Importantly, seemingly similar choices of $L(n_i)$ yield strikingly different term-weighting schemes; it would be interesting to empirically compare these new schemes against the classic IDF.

Acknowledgments. We thank Jon Kleinberg and the anonymous reviewers for helpful comments. This paper is based upon work supported in part by the National Science Foundation under grant no. IIS-0329064, a Yahoo! Research Alliance gift, and an Alfred P. Sloan Research Fellowship. Any opinions, findings, and conclusions or recommendations expressed are those of the author and do not necessarily reflect the views or official policies, either expressed or implied, of any sponsoring institutions, the U.S. government, or any other entity.

6. REFERENCES

[1] K. W. Church and W. A. Gale. Inverse document frequency (IDF): A measure of deviations from Poisson. In *Proceedings of the Third Workshop on Very Large Corpora (WVLC)*, pages 121–130, 1995.

[2] W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35(4):285–295, 1979. Reprinted in Karen Spärck Jones and Peter Willett, eds., *Readings in Information Retrieval*, Morgan Kaufmann, pp. 339–344, 1997.

[3] A. P. de Vries and T. Roelleke. Relevance information: A loss of entropy but a gain for idf? In *Proceedings of SIGIR*, pages 282–289, 2005.

[4] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *Proceedings of SIGIR*, pages 49–56, 2004.

[5] W. R. Greiff. A theory of term weighting based on exploratory data analysis. In *Proceedings of SIGIR*, pages 11–19, New York, NY, USA, 1998.

[6] D. Harman. The history of IDF and its influences on IR and other fields. In *Charting a New Course: Natural Language Processing and Information Retrieval: Essays in Honour of Karen Spärck Jones*, pages 69–79. Springer, 2005.

[7] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*, chapter 11 (Probabilistic information retrieval). Cambridge University Press, 2007. Draft of April 28.

[8] K. Papineni. Why inverse document frequency? In *Proceedings of the NAACL*, pages 1–8, 1995.

[9] S. E. Robertson. Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520, 2004.

[10] S. E. Robertson and K. Spärck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[11] S. E. Robertson and S. Walker. On relevance weights with little relevance information. In *Proceedings of SIGIR*, pages 16–24, 1997.

[12] K. Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[13] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, second edition, 1999.

[14] S. K. M. Wong and Y. Y. Yao. A note on inverse document frequency weighting scheme [sic]. Technical Report TR-89-990, Cornell University, Ithaca, NY, USA, 1989.